bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Low base-substitution mutation rate in the ciliate *Tetrahymena thermophila*

Hongan Long[1,2,a], David J. Winter[3,a,*], Allan Y-C. Chang[1], Way Sung[2], Steven H. Wu[3], Mariel Balboa[1], Ricardo B. R. Azevedo[1], Reed A. Cartwright[3,4], Michael Lynch[2], Rebecca A. Zufall[1]

1. Department of Biology and Biochemistry, University of Houston, Houston, Texas USA 77204-5001

2. Department of Biology, Indiana University, Bloomington, Indiana USA 47405

3. The Biodesign Institute, Arizona State University, Tempe, Arizona USA 85287-5301

4. School of Life Sciences, Arizona State University, Tempe, Arizona USA 85287-5301

a. These authors contributed equally to this work.

* To whom correspondence may be addressed. Email: djwinter@asu.edu

## ABSTRACT

Mutation is the ultimate source of all genetic variation and is, therefore, central to evolutionary change. Previous work on the ciliate *Paramecium tetraurelia* concluded that the presence of a transcriptionally silent germline genome has caused the evolution of a low base-substitution mutation rate in the germline genome of that ciliate. Here, we use mutation accumulation (MA) lines of the ciliate *Tetrahymena thermophila* to test the generality of this result. We find that both ciliates have similar base-substitution mutation rates in their germline genomes. The estimated base-substitution mutation rate cannot account for the observed fitness decline of the MA lines of *T. thermophila*, suggesting that the fitness decline may be caused by other factors.

Key words: microbial eukaryote, neutral evolution, effective population size,, population genetics, mutation rate

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# INTRODUCTION

Mutation is the ultimate source of all genetic variation (Baer 2008), and the rate, molecular spectrum, and phenotypic consequences of new mutations are all important drivers of biological processes such as adaptation, the evolution of sex, maintenance of genetic variation, ageing, and cancer. However, because mutations are rare, detecting them is difficult, often requiring the comparison of genotypes that have diverged from a common ancestor by at least hundreds or thousands of generations. Further, interpreting the results of such comparisons is complicated by the fact that mutations are frequently eliminated by selection before they can be studied.

Mutation accumulation (MA) is a standard method for studying mutations experimentally. In a typical MA experiment, many isolated inbred or clonal lines are passed repeatedly through bottlenecks. This reduces the effective population size and lessens the effect of selection, allowing all but the most deleterious mutations to drift to fixation (Lynch and Walsh 1998; Mukai 1964). The genome-wide mutation rate and mutational spectrum can then be estimated by comparing the genomes of MA lines and their ancestors. Such direct estimates of mutational parameters have become increasingly available for a number of model organisms (Denver et al. 2009; Keightley 2009; Keightley et al. 2014; Lee et al. 2012; Lind and Andersson 2008; Lynch et al. 2008; Ness et al. 2012; Ossowski et al. 2010; Sung et al. 2012b; Zhu et al. 2014), although a narrow phylogenetic sampling of species still limits our ability to understand how mutation rates and patterns have evolved and, in turn, have influenced evolution across the Tree of Life.

Microbial eukaryotes are an extraordinarily diverse group, containing many evolutionarily distant lineages, some of which have unusual life-histories and genome features (Katz and Bhattacharya 2006). Microbial eukaryotes are, however, often unsuitable for use in mutational studies because they are difficult to culture in the lab, especially at the small population sizes required for MA experiments. In addition, most eukaryotic microbes lack genomic resources (e.g., finished annotated reference genome). The high proportion of repeated and duplicated DNA that reside in the large genomes of many of these

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

species makes it difficult to produce reference sequences, contributing to the incomplete nature of many existing resources. These barriers have limited mutation-rate studies in microbial eukaryotes to *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Paramecium tetraurelia*, *Dictyostelium discoideum*, and *Chlamydomonas reinhardtii* (Farlow et al. 2015; Lynch et al. 2008; Ness et al. 2012; Saxer et al. 2012; Sung et al. 2012b; Zhu et al. 2014) (Figure 1).

The ciliated unicellular eukaryote *Tetrahymena thermophila* is particularly well suited to MA experiments. Like all ciliates, individuals from this species have distinct germline and somatic copies of their nuclear genome. During asexual growth, the contents of the germline genome are duplicated mitotically but never expressed. Unlike most other ciliates (including *P. tetraurelia*, which senesces in the absence of periodic mating or autogamy), *T. thermophila* can be propagated this way indefinitely. Thus, mutations can accumulate in the germline genome over thousands of asexual generations while never being subjected to natural selection. Long et al. (2013) confirmed that *T. thermophila* can be maintained in MA lines, growing asexually for 1000 generations, and studied the fitness effects of newly arising mutations but did not directly measure the mutation rate.

The only existing MA experiment from a ciliate *Paramecium tetraurelia* (Sung, et al. 2012b) yielded the lowest known nucleotide mutation rate. Measurement of the mutation rate of *T. thermophila* will help reveal whether a low mutation rate is a general feature of ciliates. In addition, natural populations of *T. thermophila* have been the focus of population-genetic studies (Zufall, et al. 2013, Katz et al. 2006), so mutational parameters estimated from MA experiments can be related to population and evolutionary processes.

Although the life history of *T. thermophila* is ideal for MA experiments, some features of its genome complicate typical computational approaches to detecting mutations from short-read sequencing. In particular, the genome is extremely AT-rich (~78% AT) and contains low-complexity and repetitive elements. These features, combined with an incomplete reference genome (Eisen et al. 2006) make

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mapping sequencing reads to the reference difficult, producing false positives using naive mutation detection methods. To overcome these difficulties, we have both used an existing pipeline and developed a novel mutation-detection approach, which directly models the design of an MA experiment and accommodates the noise introduced by mismapped reads.

Here we expand the work presented by Long et al. (2013) by directly estimating the nucleotide mutation rate in *T. thermophila*. Our results are consistent with the exceptionally low rate estimated for *P. tetraurelia*, indicating that low germline mutation rates may be a feature of ciliates generally. We also use our estimated rate to calculate the effective population size of *T. thermophila* in the wild. Our results establish that direct estimates of the mutation rate from sequence data are possible for *T. thermophila*, setting the scene for longer or larger MA experiments that will be required to estimate the mutational spectrum of a species with such a low mutation rate.

# RESULTS

## Putative mutations and validation

To estimate the micronuclear mutation rate, we sequenced the whole macronuclear genomes of 10 homozygous genomic-exclusion lines, each derived from a separate *T. thermophila* line that had undergone MA for approximately 1000 generations. We identified 93 sites for which there was some evidence of a mutation in at least one lineage. On closer inspection of the data underpinning these putative mutations, we found an unusual pattern — more than half of the apparent mutations were from lines M47 and M51, and in many cases reads containing the apparent-mutant allele from one of these lines were also sequenced from the other line (but absent or very rare in all other lines).

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To investigate this pattern further we analyzed the frequency of non-reference bases in all samples across the whole genome (Supplementary Data). These analyses demonstrated that M47 and M51 differ from all other lines in the frequency of non-reference bases and in patterns of sequencing coverage. We do not know what caused this pattern. It is possible that some cellular process occurred in these lines but not others (e.g., the incorporation of sequences usually restricted to the micronucleus, or the inclusion of DNA from the star strain during genomic exclusion). It is extremely unlikely that M47 and M51 independently accrued more shared mutations than independent mutations during our MA experiment. For this reason we have excluded these lines from all subsequent analyses.

Forty one putative mutations remained after lines 47 and 51 were removed. We attempted to validate each of these mutations using Sanger sequencing. Only 4 of these mutations were validated, with the remaining sites where either shown to be false positives (11 sites) or failing to generate PCR amplicons or clean sequencing traces (26 sites). Closer inspection of the data underpinning both the false positive and inconclusive mutations showed these sites to have unusually low sequencing coverage, low mapping quality and to be subject to strand bias. All of these properties are associated with difficult to map regions of genomes, and are known to generate false positive variant calls (Li 2014). For this reason, we re-ran our probabilistic mutation caller using stricter parameter values and excluding sites that did not have at least 3 sequencing reads supporting a mutation in both the forward and reverse orientation. None of the inconclusive or false positive sites were called as mutations in this analysis, which also detected an additional mutation that was confirmed by Sanger sequencing. Thus, we detected a total of 5 mutations across 8 MA lines, with no line having more than one confirmed mutation (Table 1). Two of these mutations produce non-synonymous substitutions, while two others fall in genes but do not affect the protein sequence, and the final mutation is in an intergenic region.

## Mutation rate

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Given the number of callable sites, the 5 mutations that we detected yield a base-substitution mutation rate estimate of $7.61 \times 10^{-12}$ (95% CI = $[0.691 \times 10^{-12}, 14.529 \times 10^{-12}]$). This point-estimate is ~1/3 of the rate reported for *P. tetraurelia*, although the 95% confidence intervals of both estimates overlap (Figure 1).

If our estimate of the base-substitution mutation rate holds for the portions of the genome from which we did not have sufficient power to detect mutations (see Methods), then we estimate that we have failed to detect an additional 0.87 mutations in the MAC genome, giving rise to a genome-wide base-substitution mutation rate of 0.0008 mutations per haploid genome per generation (95% CI = [0.00007, 0.0015]).

## DISCUSSION

We have used whole genome sequencing and a novel mutation-detection approach to estimate the base-substitution mutation rate of *T. thermophila* from 8 MA lines (Long et al. 2013), and obtained an estimate of $7.61 \times 10^{-12}$ base-substitution mutations per site per generation. This is the lowest estimate of mutation rate from a eukaryote (see Figure 1, and Sung *et al.* (2012b) for surveys of mutation-rate estimates), although it is not statistically significantly different from those of either the social amoeba *Dictyostelium discoideum* or the ciliate *P. tetraurelia* (Lynch and Conery 2003; Sung et al. 2012b). The fact that the two lowest mutation rates have been recorded in ciliates supports the hypothesis that ciliates in general have low germline mutation rates (Sung *et al.* 2012b).

The unusual genome structure and life history of ciliates may explain their low mutation rates. Sung et al. (2012a) argue that mutation rates are minimized to the extent made possible by the power of natural selection — the “drift barrier” hypothesis (Sung et al. 2012a). Because the mutations accumulating in the germline genome during asexual reproduction in ciliates are not exposed to natural selection, the mutation rate per selective event is equal to the number of cell divisions between rounds of sexual reproduction

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

multiplied by the germline per-division mutation rate. Thus, the low per-division mutation rates reported for ciliates may have evolved naturally as a consequence of a prolonged time for germline turnover.

Unlike *P. tetraurelia*, *T. thermophila* does not undergo senescence in the absence of sex, and we lack a good estimate for the frequency of sexual reproduction in natural populations (Doerder et al. 1995). Therefore, we cannot put an upper bound on the number of asexual generations between conjugation events. However, we can estimate a lower bound because cells arising from sexual reproduction enter a period of immaturity lasting ~50-100 divisions (Lynn and Doerder 2012). We know that the germline genome divides at least this many times without opportunity for selection on any newly acquired mutations. Using the immaturity period as a proxy for frequency of sex gives an estimate of the base-substitution mutation rate per conjugation event that is much closer to that of other eukaryotes per round of DNA replication (Sung et al. 2012b).

The unusual genome structure of ciliates may also present a novel test of the drift-barrier hypothesis of mutation rate evolution (Sung et al. 2012a). If the mutation rates of the germline and somatic nuclei can evolve independently then we would expect the somatic mutation rate to be higher (i.e., more similar to the mutation rates of other eukaryotes) because somatic mutations are exposed to selection after each cell division. At present, there are no estimates of the somatic mutation rate of ciliates.

Most mutations with effects on fitness are deleterious, so the accumulation of mutations in the absence of selection is expected to lead to a reduction in organismal fitness (Bateman 1959; Halligan and Keightley 2009; Mukai 1964; Muller 1928). Indeed, each of the MA lines from which we detected base substitutions experienced substantial fitness losses over the course of our experiment (Long et al. 2013). It is difficult, however, to see how the mutations that we have detected could explain these declines in fitness. We did not detect any base-substitution mutations in the line with largest observed loss in fitness (M50, w=0.38), while those lines with non-synonymous mutations (which we might expect to have the most severe fitness consequences) do not have especially low fitness values (Table 1). It is unlikely that

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the fitness losses observed in these MA lines can be explained by other undetected single-base substitutions, as our mutation calling method had power to detect mutations in an average of 86.1% of the genome (Table 1). Rather, it seems likely the fitness of these lines is determined in part by indels and other structural variants that we did not include in this study. Future work could explicitly test this prediction. Alternatively, non-Mendelian patterns of inheritance could complicate the relationship between mutations and fitness measurements. For example, the fitness of an individual line may not only be influenced by mutations but also by epigenetic processes, such as cortical inheritance (Sonneborn 1963) or small RNA guided genome rearrangement (Mochizuki and Gorovsky 2004).

Our mutation rate estimate also allows us to update previous estimates of the effective population size of *T. thermophila*. If we assume that silent sites in protein-coding genes are effectively neutral and under drift-mutation equilibrium, the population-level heterozygosity at silent sites ($\pi_s$) equals $4N_e\mu$, where $N_e$ is the effective population size, and $\mu$ is mutation rate per site per generation. Using the estimate $N_e \times \mu = 8 \times 10^{-4}$ reported by Katz et al. (2006), our $N_e$ estimate for *T. thermophila* is $1.12 \times 10^8$, which is approximately equal to that of *P. tetraurelia* ($N_e = 1.24 \times 10^8$; (Sung et al. 2012b)). These estimates may seem surprising given the observations that *P. tetraurelia* is cosmopolitan and regularly isolated from different continents (Catania et al. 2009), while *T. thermophila* has a distribution limited to the eastern United States (Zufall et al. 2013). However, these estimates may be influenced by the fact that *T. thermophila* populations have significant population structure (Zufall et al. 2013) and the combination of facultative sexuality and the unusual ciliate genome structure may result in extended persistence of deleterious alleles (Morgens et al. 2014). It is also possible that *P. tetraurelia* does indeed have a larger census population size than *T. thermophila*, but this larger population size is not reflected in the genetic diversity of these species due to the increased power of natural selection to constrain diversity at linked sites in larger populations (Corbett-Detig et al. 2015).

In this study we have established that it is possible to detect mutations in *T. thermophila* MA lines through short-read sequencing, and thus to directly study the nature of mutation in this model organism.

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Although we were able to show that *T. thermophila* shares a low mutation rate with *P. tetraurelia* (the only other ciliate for which a mutation rate has been directly estimated), there is still much to learn about mutation in this species. In particular, the small number of mutations accumulated over this experiment prevents us from analyzing the spectrum of mutations arising in *T. thermophila* and determining the influence of mutational biases on genome evolution. Similarly, the few mutations that we detect seem inadequate to explain the observed losses of fitness during MA. Future studies using more MA lines and longer periods of MA, and detecting indels and other structural variants accrued during MA will be needed to fully understand the effects of mutation in *T. thermophila*.

# MATERIALS AND METHODS

## Cell lines

The 10 evolved cell lines that were used in this study were generated from 10 parental MA lines (Table S1). These lines were established from a single cell of the strain SB210 and described in detail in Long et al. (2013). Briefly, the 10 MA lines were cultured in the rich SSP medium in test tubes (Gorovsky et al. 1975) and experienced ~50 single-cell bottlenecks and 1000 cell divisions, except for M28, which was bottlenecked for 10 times and passed ~200 cell divisions. Genomic exclusion lines were then produced by two rounds of crossing between the MA lines (mating type VI) and a germline-dysfunctional B* strain (mating type VII) (Bruns and Cassidy-Hanley 1999). 19 independent genomic exclusion operations were also done on the progenitor line M0 to express all possible ancestral genotypes out, and all these 19 independent M0 genomic exclusion lines' DNA were pooled before library construction. This was used for filtering out false positives originated from progenitor heterozygosity.

## Whole genome sequencing

DNA libraries with insert size ~350 bp and Illumina paired-end sequencing were done in the DNASU core facility at the Biodesign Institute at Arizona State University and the Hubbard Center for Genome Studies, University of New Hampshire. The mean sequencing depth is ~47×, with >90% of the sites in the

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genome covered in all the sequenced lines (Table S1). Sequencing reads are available from the NCBI's SRA database under a BioProject with accession number PRJNA285268.

## Base-substitution analysis

In order to avoid false negatives that might not be detected by a single approach. We used two independent approaches to call point-mutations: (1) a probabilistic approach that adapts methods designed for family-based data to the design of MA experiments (Cartwright et al. 2012); (2) a consensus approach which has been successfully used in other systems (Sung et al. 2012b). Our list of candidates was generated by the union of calls from these two methods.

*Probabilistic approach using accuMUlate*

We developed a probabilistic approach to detecting mutations from genomic alignments. The challenge of identifying mutations from such data is best described as a hidden-data problem in which the observed data, *R*, contains the set of reads mapped to a site, while the hidden data, *H*, contains ancestral and descendent genotypes, meioses, locations of mutations, and locations of sequencing errors. Each unique combination of hidden states in the model represents a different potential history during our experiment. We developed a probabilistic approach to detecting mutations from our data by adapting hidden-data methods designed for family-based studies (Cartwright et al. 2012) to MA experiments.

Our approach uses a model to calculate the probability that a given site contains a *de novo* mutation, using our sequencing data as the only observed input:

$$P(\text{mutation}|R;\Theta) = \frac{P(\text{mutation}, R;\Theta)}{P(R;\Theta)}$$

Where $\Theta$ represents the model parameters and contains

- $\theta$ The approximate proportion of sites in the ancestor which are heterozygous
- $\varphi_A$ The over-dispersion parameter for sequencing of the ancestor (described below)
- $\varphi_D$ The over-dispersion parameter for sequencing of the descendant lines (described below)
- $\pi$ A vector representing the proportion of each nucleotide present in the ancestral genome
- $\mu$ The experiment-long mutation rate per site

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- $\varepsilon$ The per-site sequencing error rate

The numerator and denominator in the first equation are marginal probabilities, and in order to calculate them from the full data, we sum over all possible histories:

$$P(\text{mutation}|R;\Theta) = \frac{\sum_H P(\text{mutation}, R, H; \Theta)}{\sum_H P(R, H; \Theta)}$$

$$= \frac{\sum_H P(\text{mutation} \mid H)\, P(R, H; \Theta)}{\sum_H P(R, H; \Theta)}$$

Because the presence or absence of mutation in a line is itself a part of any given state *H*, this approach amounts to finding the sum of the probabilities of those histories that contain a mutation. Therefore, we only need to determine how to calculate $P(R, H; \Theta)$, the probability of the full data given the model parameters. For a given history this probability is found via

$$P(R, H|\Theta) = P(G_A \mid \theta, \pi) \cdot P(R_A|\ G_A;\ \varphi_A, \varepsilon) \cdot \prod_i^n [P(G_i \mid\ \pi, \mu) \cdot P(R_i \mid G_i;\ \varphi_D, \varepsilon)]$$

Where $G_A$ is the ancestral genotype in this history, $R_A$ are the reads generated from the ancestral strain and $R_i$ and $G_i$ are the reads and genotype of the *i*-th of *n* descendant lines. As this equation is for a specific history, the terms $P(R_i \mid G_i;\ \varphi_D, \epsilon)$ and $P(R_A|\ G_A;\ \varphi_A, \epsilon)$ are the probability of a set of base calls given a genotype and model parameters (sometimes called the genotype likelihood). We calculate these genotype likelihoods using the Dirichlet-Multinomial distribution. For example, let $\boldsymbol{R}$ be a vector which contains the number of A, C, G and T bases mapped to a given site in a given sample, the probability mass function for any genotype is given by

$$P(\boldsymbol{R};\ \varphi, \boldsymbol{p}) = \binom{N}{\boldsymbol{R}} \frac{\Gamma(1-\varphi)}{\Gamma\big(1+\varphi(N-1)\big)} \prod_i \frac{\Gamma\big(\boldsymbol{p}_i + \varphi(\boldsymbol{R}_i - \boldsymbol{p}_i)\big)}{\Gamma(\boldsymbol{p}_i - \varphi \boldsymbol{p}_i)}$$

where *N* is the total number of reads, Γ is the gamma function, and $\varphi$ is the expected positive correlation between pairs of base calls (and thus a measure of over-dispersion relative to a standard multinomial

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

distribution). To calculate the likelihood of a particular genotype we set the elements of $\boldsymbol{p}$ to reflect alleles in that genotype and the probability of sequencing error. So, to calculate the likelihood for a diploid genotype $g_{jk}$ we set the values of $\boldsymbol{p}$ as follows

$$\boldsymbol{p}_i = \begin{cases} 1-\varepsilon, & \text{if } i = g_j = g_k \\ \frac{1}{2}-\frac{\varepsilon}{3}, & \text{if } i = g_j \neq g_k \text{ or } i = g_k \neq g_j \\ \frac{\varepsilon}{3}, & \text{otherwise} \end{cases}$$

The remaining terms in equation describing the probability of a given history, $P(G_A \mid \theta, \pi)$ and $P(G_i \mid \pi, \mu)$, represent the prior probabilities of ancestral and specific descendant genotypes respectively. We use the Dirichlet-Multinomial distribution to calculate the prior probability of an ancestral genotype, taking into account the expected heterozygosity and nucleotide composition of the ancestral genome. Specifically, we calculate the probability of a given ancestral genotype $a_{jk}$ using the following equation with $\varphi = 1/(1+\theta)$ and $\boldsymbol{Z}$ as a vector of length 4 with values corresponding to the number of A, C, G, and T alleles in the genotype.

$$P(a_{jk}|\theta, \boldsymbol{\pi}) = \binom{N}{\boldsymbol{Z}} \frac{\Gamma(1-\varphi)}{\Gamma(1+\varphi(N-1))} \prod_i \frac{\Gamma(p_i + \varphi(\boldsymbol{Z}_i - \boldsymbol{\pi}_i))}{\Gamma(\boldsymbol{\pi}_i - \varphi\boldsymbol{\pi}_i)}$$

We use the (Felsenstein 1981) model of nucleotide substitution to calculate the prior probability of specific descendant genotype $d_l$ arising from the ancestral genotype in a given history. This model incorporates equilibrium nucleotide frequencies, allowing us to incorporate the extreme AT-bias present in the *T. thermophila* genome in our model.

The approach we describe can easily be adapted to find the sum of the probabilities of all histories where only one mutation occurs. We also calculated this value, because it is very unlikely that multiple lines will accrue mutations in the same site during an MA experiment, but quite possible that systematic errors in sequencing and mapping reads to reference will generate a mutation-like pattern in multiple samples.

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This model is implemented in a C++ program called accuMUlate, which makes use of the Bamtools (Barnett et al. 2011) library. The source code is available under an MIT license from https://github.com/dwinter/accumulate the specific version of the code used in these analyses is archived at http://dx.doi.org/10.5281/zenodo.19942 . We ran our model on a genomic alignment produced by using Bowtie version 2.1.0 (Langmead and Salzberg 2012) to map reads to the December 2011 release of the *T. thermophila* macronuclear genome from the *Tetrahymena* Genome Database (Stover et al. 2006). One site in the reference contained a gap character, which we removed since our reads indicated that it was an artifact. We processed the resulting alignments to remove sequencing and mapping artifacts that could lead to false-positive mutation calls. In particular, we identified and marked duplicate reads using the MarkDuplicates tool from Picard 1.106 (http://pricard.sourceforge.net) and performed local realignment around potential indels using GATK 3.2 (DePristo et al. 2011; McKenna et al. 2010). We adjusted raw base quality scores by running GATK's BaseRecalibrator tool, using a set of putative single nucleotide variants detected with SAMtools mpileup as input (Li et al. 2009).

The putative mutations from this approach were preliminarily identified by running accuMUlate to identify sites with a mutation probability > 0.1 with a relatively lenient set of parameter-values: $\varphi_A = 0.001, \varphi_D = 0.001, \varepsilon = 0.01, \mu = 10^{-8}$ and only considering reads with mapping-quality ≥ 13. The validation phase showed that false-positive mutations were frequently associated with poorly-mapped reads, low coverage regions surrounding deletions with respect to the reference genome, or the presence of rare bases in all samples. Thus, we re-ran the accuMUlate model, excluding all reads with a mapping quality < 25, and using the more conservative over-dispersion parameters $\varphi_A = 0.03$ and $\varphi_D = 0.01$. In addition, we filtered out putative mutations that were not supported by at least 3 reads in both forward and reverse orientation.

*Consensus approach*

Putative mutations were called if one individual line is different from the consensus of all the remaining lines, after filtering out mismapping/library PCR/sequencing errors and reads with low quality and

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mapping scores (< 20), with two mapping programs BWA 0.7.10 (Li and Durbin 2009) and novoalign (V2.08.01; NOVOCRAFT Inc) (Sung et al. 2012a). Analyzed sites passing the filters were used as the denominator to calculate mutation rate. This approach has been applied in a wide variety of prokaryotic and eukaryotic organisms and repeatedly verified with Sanger sequencing (Denver et al. 2009; Lee et al. 2012; Long et al. 2015; Ossowski et al. 2010; Sung et al. 2015).

## Validation of putative mutations

The validity of a subset of putative mutations was tested by Sanger sequencing. All mutations identified by either the accuMUlate or consensus approach were tested with suitable primers up to 500 bp away from the mutation site using the default parameters of Primer3 (Koressaar and Remm 2007; Untergrasser et al. 2012) as implemented in Geneious (http://www.geneious.com, (Kearse et al. 2012)). Successful PCRs were purified and directly sequenced at Lone Star Labs (Houston, TX).

## Mutation rate calculations

Our probabilistic approach to mutation detection also provides a straightforward means to calculate the number of sites at which we could have detected a mutation if one was present, and thus the correct denominator to use for mutation rate calculations. We did this by re-calculating our mutation probabilities having first introduced simulated mutations in one sample by shuffling the vector of read-counts for that sample. This procedure was repeated for every site in the reference genome, shuffling the read counts from each descendent separately. A site was treated as missing from a sample if the mutation probability calculated from shuffled read-counts was < 0.1. The number of callable sites detected using this approach for each line is given in Table 1.

We calculated the mutation rate by summing the number of validated mutations across MA lines, and then dividing it by the product of analyzed sites and generations in each MA line: $\hat{\mu} = \sum_i n_i / L\,T$ Assuming that the number of mutations in each line follows a Poisson distribution (but not necessarily the same distribution), the standard error for our estimate of mutation rate was estimated as $\mathrm{SE}(\hat{\mu}) = \sqrt{\hat{\mu}/LT}$ and a 95% confidence interval was constructed as $\hat{\mu} \pm 1.96\, SE(\hat{\mu})$.

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To calculate genomic mutation rates we assumed a haploid genome size of 104 Mb (Eisen et al. 2006).

## Annotation of mutations

We annotated the functional context of identified mutations, using snpEff (Cingolani et al. 2012) and the December 2011 release of the *T. thermophila* macronuclear genome annotation file from the *Tetrahymena* Genome Database.

ACKNOWLEDGEMENTS

We thank Kristen Dimond, Robert Coyne, Tom Doak, Kale Dai, and Rachel Schwartz for technical help. This study is funded by NIH R01GM101352 (RAZ, RBRA, RAC) and Multidisciplinary University Research Initiative award W911NF-09-1-0444 (ML) from the US Army Research Office, NIH grant R01GM101672 (ML) and National Science Foundation Grant MCB-1050161 (ML).

REFERENCES

Baer CF 2008. Does mutation rate depend on itself? PLoS Biol. 6: 233–235.
Bateman AJ 1959. The viability of near-normal irradiated chromosomes. Int. J. Radiat. Biol. 1: 170–180.
Bruns P, Cassidy-Hanley D 1999. Methods for genetic analysis. Methods Cell Biol. 62: 229–240.
Cartwright RA, Hussin J, Keebler JEM, Stone EA, Awadalla P 2012. A family-based probabilistic method for capturing de novo mutations from high-throughput short-read sequencing data. Stat. Appl. Genet. Molec. 11: Article 6.
Catania F, Wurmser F, Potekhin AA, Przybo E, Lynch M 2009. Genetic diversity in the *Paramecium aurelia* species complex. Mol Biol Evol 26: 421–431.
Corbett-Detig RB, Hartl DL, Sackton TB 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13: e1002112.
Denver DR, et al. 2009. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. Proc Natl Acad Sci USA 106: 16310–16314.
DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491–498.
Doerder FP, Gates MA, Eberhardt FP, Arslanyolu M 1995. High frequency of sex and equal frequencies of mating types in natural populations of the ciliate *Tetrahymena thermophila*. Proc. Natl. Acad. Sci. USA 92: 8715–8718.
Eisen J, Coyne R, Wu M, Wu D, Thiagarajan M 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 4: 1620–1642.
Farlow A, et al. 2015. The spontaneous mutation rate in the fission yeast *Schizosaccharomyces pombe*. Genetics. doi: 10.1534/genetics.115.177329
Felsenstein J 1981. Evolutionary trees from DNA-sequences - a maximum-likelihood approach. J. Mol. Evol. 17: 368–376. doi: Doi 10.1007/Bf01734359

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gorovsky M, Yao M, Keevert J, Pleger G 1975. Isolation of micro- and macronuclei of *Tetrahymena pyriformis*. Methods Cell Biol. 9: 311–327.

Halligan DL, Keightley PD 2009. Spontaneous mutation accumulation studies in evolutionary genetics. Annu. Rev. Ecol. Evol. Syst. 40: 151–172.

Hinchliff C, et al. 2014. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. bioRxiv. doi: http://dx.doi.org/10.1101/012260

Katz LA, Bhattacharya D. 2006. Genomics and evolution of microbial eukaryotes. Oxford, UK: Oxford University Press.

Katz LA, Snoeyenbos-West O, Doerder FP 2006. Patterns of protein evolution in *Tetrahymena thermophila*: implications for estimates of effective population size. Mol Biol Evol 23: 608–614.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649.

Keightley P, Trivedi U, Thomson M, Oliver F, Kumar S et al. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res 19: 1195–1201.

Keightley PD, Ness RW, Halligan DL, Haddrill PR 2014. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. Genetics 196: 313–320.

Koressaar T, Remm M 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23: 1289–1291.

Langmead B, Salzberg SL 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.

Lee H, Popodi EM, Tang H, Foster PL 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc Natl Acad Sci USA 109: E2774–E2783.

Li H 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 30: 2843–2851.

Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754–1760.

Li H, et al. 2009. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078–2079.

Lind PA, Andersson DI 2008. Whole-genome mutational biases in bacteria. Proc Natl Acad Sci USA 105: 17878–17883.

Long H, Paixao T, Azevedo RBR, Zufall RA 2013. Accumulation of spontaneous mutations in the ciliate *Tetrahymena thermophila*. Genetics 195: 527–540.

Long H, et al. 2015. Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair (MMR) deficient *Pseudomonas fluorescens* ATCC948. Genome Biol Evol 7: 262–271.

Lynch M, Conery JS 2003. The origins of genome complexity. Science 302: 1401–1404.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. USA 105: 9272–9277.

Lynch M, Walsh B. 1998. Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer.

Lynn DH, Doerder FP 2012. The life and times of *Tetrahymena*. Methods Cell Biol. 109: 9–27.

McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

Mochizuki K, Gorovsky MA 2004. Conjugation-specific small RNAs in *Tetrahymena* have predicted properties of scan (scn) RNAs involved in genome rearrangement. Genes Dev. 18: 2068–2073.

Morgens DW, Stutz TC, Cavalcanti AR 2014. Novel population genetics in ciliates due to life cycle and nuclear dimorphism. Mol. Biol. Evol. 31: 2084–2093.

Mukai T 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. Genetics 50: 1–19.

Muller HJ 1928. The measurement of gene mutation rate in *Drosophila*, its high variability, and its dependence upon temperature. Genetics 13: 279–357.

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Ness RW, Morgan AD, Colegrave N, Keightley PD 2012. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. Genetics 192: 1447–1454.

Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD 2015. Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. Genome Res. doi: 10.1101/gr.191494.115

Ossowski S, et al. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327: 92–94.

Saxer G, et al. 2012. Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba *Dictyostelium discoideum*. PLoS One 7: e46759.

Sonneborn TM. 1963. Does preformed cell structure play an essential role in cell heredity. In: Allen JM, editor. The nature of biological diversity. New York: McGraw-Hill.

Sonneborn TM 1954. The relation of autogamy to senescence and rejuvenescence in *Paramecium aurelia*. J. Eukaryot. Microbiol. 1: 38–53.

Stover NA, et al. 2006. *Tetrahymena* Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research. Nucleic Acids Res. 34: D500–D503.

Sung W, et al. 2015. Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. Mol Biol Evol. 32: 1672–1683.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M 2012a. Drift-barrier hypothesis and mutation rate evolution. Proc Natl Acad Sci USA 109: 18488–18492.

Sung W, et al. 2012b. Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. Proc Natl Acad Sci USA 109: 19339–19344.

Untergrasser A, et al. 2012. Primers—new capabilities and interfaces. Nucl Acids Res 40: e115.

Zhu YO, Siegal ML, Hall DW, Petrov DA 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA 111: E2310–E2318.

Zufall RA, Dimond KL, Doerder FP 2013. Restricted distribution and limited gene flow in the model ciliate *Tetrahymena thermophila*. Mol Ecol 22: 1081–1091.

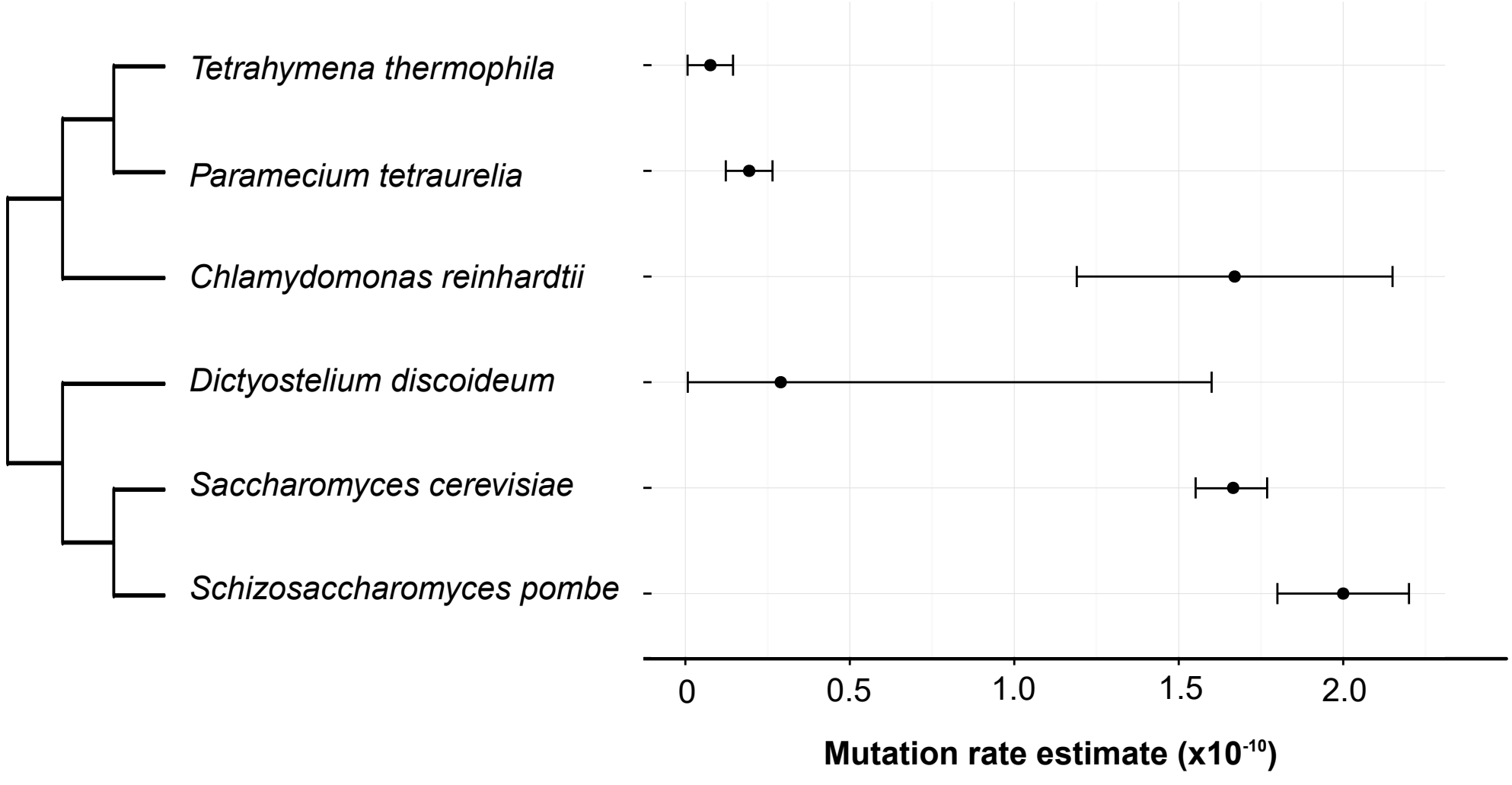
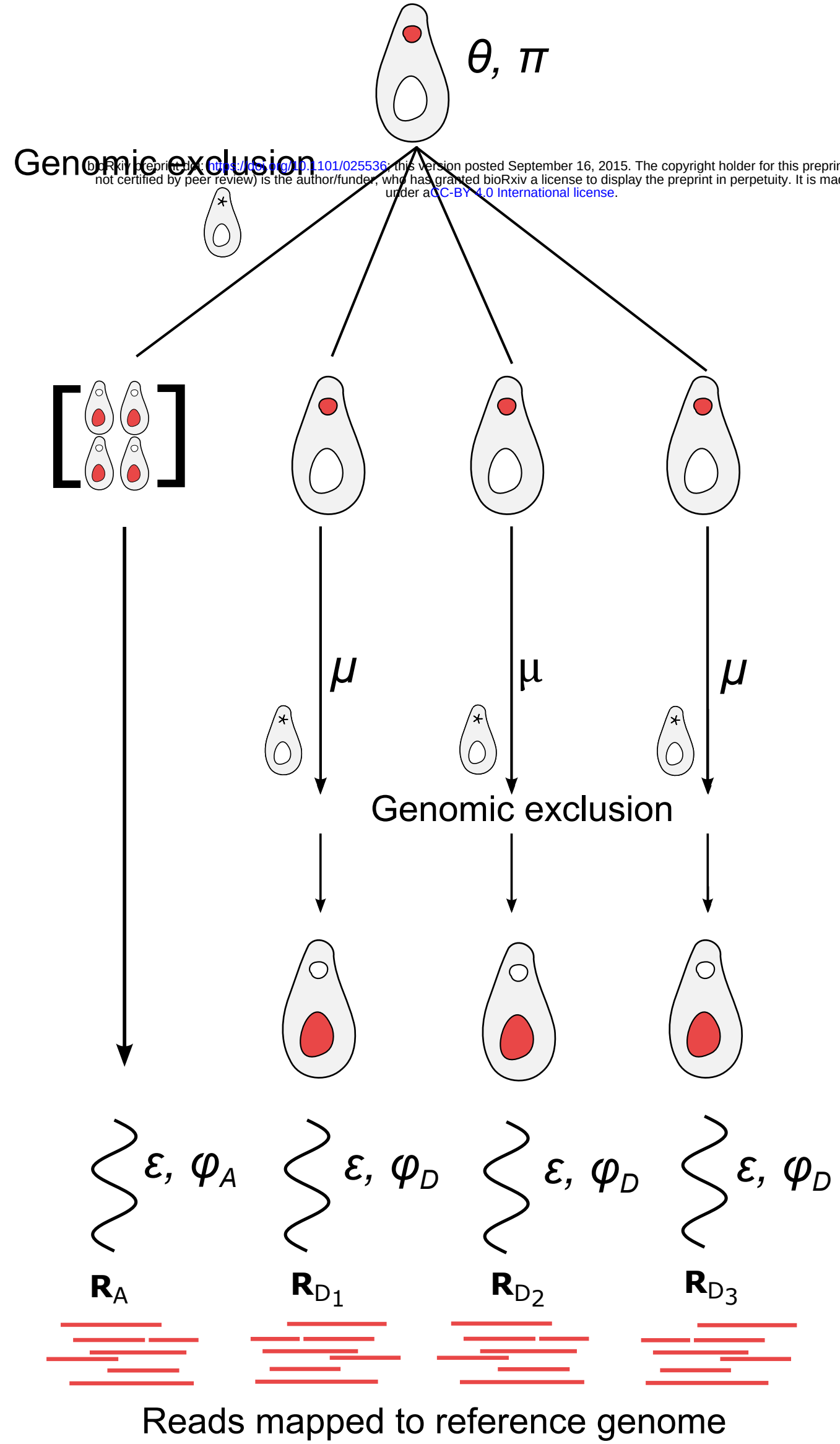
**Figure 1 Mutation rate estimates for unicellular eukaryotes.**

Base-substitution mutation rates per nucleotide per generation estimated for different unicellular eukaryotes: *T. thermophila* (this paper), *P. tetraurelia* (Sung et al. 2012b), *C. reinhardtii* (Ness et al. 2015), *D. discoideum* (Saxer et al. 2012) , *Sa. cerevisiae* (Zhu et al. 2014), and *Sc. pombe* (Farlow et al. 2015). Error bars are 95% confidence intervals. The phylogenetic tree was retrieved from the Open Tree of Life (Hinchliff et al. 2014); branch lengths are arbitrary.

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**Figure 2 Experimental design in relation to model parameters.**

A complete description of the experiment is presented in Long et al. (2013). Here, we describe how the experiment relates to the parameters used in our mutation-calling model. Specifically, the ancestral line with average heterozygosity $\theta$ and genome-wide nucleotide frequencies $\boldsymbol{\pi}$ is used to generate a set of MA lines. Each of these lines accumulate mutations at a rate $\mu$ per nucleotide per generation for 1000 generations. Genomic exclusion, an auto-diploidization process, is used to generate lines with macronuclei representing one haploid-copy of each MA line (and multiple copies of the ancestral line, in order to detect ancestral heterozygosity). These lines that have undergone genomic exclusion are then sequenced with a sequencing error rate of $\varepsilon$ and over-dispersion caused by library preparation and other correlated errors modeled as $\varphi_A$ and $\varphi_D$ for ancestral and descendant lines respectively. A full description of this model and its parameters is given in the subsection of the Materials and Methods labeled "Probabilistic approach using accuMUlate".


Tetrahymena thermophila
Paramecium tetraurelia
Chlamydomonas reinhardtii
Dictyostelium discoideum
Saccharomyces cerevisiae
Schizosaccharomyces pombe
0
0.5
1.0
1.5
2.0
Mutation rate estimate (x10-10)

bioRxiv preprint doi: https://doi.org/10.1101/025536; this version posted September 16, 2015. The copyright holder for this preprint (not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.